(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
15. November 2001 (15.11.2001)

PCT

(10) Internationale Veröffentlichungsnummer
WO 01/86634 A1

(51) Internationale Patentklassifikation[7]: G10L 15/06

(21) Internationales Aktenzeichen: PCT/DE01/01546

(22) Internationales Anmeldedatum:
24. April 2001 (24.04.2001)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
100 22 586.1 9. Mai 2000 (09.05.2000) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **SIEMENS AKTIENGESELLSCHAFT** [DE/DE]; Wittelsbacherplatz 2, 80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: **BUDDE, Mark** [DE/DE]; Agathastr. 8, 46240 Bottrop (DE). **SCHNEIDER, Tobias** [DE/DE]; Kranzhornstr. 7, 81825 München (DE).

(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, 80506 München (DE).

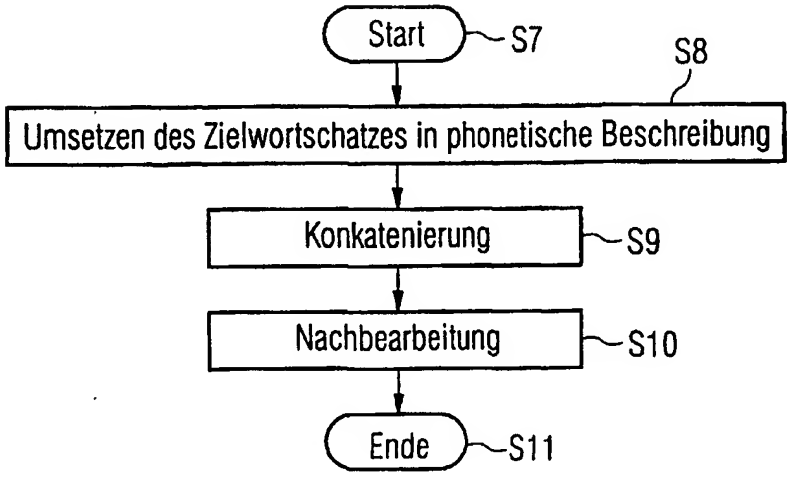
(81) Bestimmungsstaat *(national)*: US.

(84) Bestimmungsstaaten *(regional)*: europäisches Patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

[Fortsetzung auf der nächsten Seite]

(54) Title: METHOD FOR CREATING A SPEECH DATABASE FOR A TARGET VOCABULARY IN ORDER TO TRAIN A SPEECH RECOGNITION SYSTEM

(54) Bezeichnung: VERFAHREN ZUM ERZEUGEN EINER SPRACHDATENBANK FÜR EINEN ZIELWORTSCHATZ ZUM TRAINIEREN EINES SPRACHERKENNUNGSSYSTEMS



S8... .CONVERSION OF THE TARGET VOCABULARY INTO PHONETIC DESCRIPTION
S9... .CONCATENATION
S10...SUBSEQUENT PROCESSING
S11...END

(57) Abstract: According to the invention, the words of the target vocabulary are composed of segments, which consist of one or more phonemes, whereby the segments are derived from a training text that is independent from the target vocabulary. The training text can be an arbitrary generic text.

[Fortsetzung auf der nächsten Seite]

WO 01/86634 A1

Veröffentlicht:
— *mit internationalem Recherchenbericht*

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

(57) Zusammenfassung: Erfindungsgemäß werden die Worte des Zielwortschatzes aus Segmenten zusammengesetzt, die aus einem oder mehreren Phonen bestehen, wobei die Segmente von einem vom Zielwortschatz unabhängigen Trainingstextes abgeleitet sind. Der Trainingstext kann ein beliebiger allgemeiner Text sein.

Beschreibung

Verfahren zum Erzeugen einer Sprachdatenbank für einen Zielwortschatz zum Trainieren eines Spracherkennungssystems

Die Erfindung betrifft ein Verfahren zum Erzeugen einer Sprachdatenbank für einen Zielwortschatz zum Trainieren eines Spracherkennungssystems.

Spracherkennungssysteme gibt es für unterschiedlichste Anwendungen. Z.B. werden in automatischen Diktiersystemen Spracherkennungssysteme verwendet, die einen sehr umfangreichen Wortschatz erkennen können, jedoch üblicherweise benutzerspezifisch ausgebildet sind, das heißt, dass sie lediglich von einem einzigen Benutzer verwendet werden können, der das Spracherkennungssystem auf seine persönliche Aussprache trainiert hat. Automatische Vermittlungssysteme in Telefonanlagen verwenden hingegen sprecherunabhängige Spracherkennungssysteme. Diese Spracherkennungssysteme benötigen einen wesentlich geringeren Wortschatz, da bei Telefonvermittlungssystemen z.B. nur wenige unterschiedliche Wörter gesprochen werden, um sich mit einem Fernsprechteilnehmer verbinden zu lassen.

Herkömmlicherweise wurde für sprecherunabhängige Spracherkennungssysteme ein Zielwortschatz (Applikationswortschatz) festgelegt. Es werden dann Trainingstexte zusammengestellt, die überwiegend Wörter aus diesem Zielwortschatz enthalten. Diese Trainingstexte werden von Sprechern gesprochen und über ein Mikrofon aufgezeichnet. Üblicherweise lässt man einen solchen Trainingstext von 100 bis 5000 Sprechern sprechen. Die gesprochenen Texte liegen somit als elektrische Sprachsignale vor. Die zu sprechenden Texte werden auch in ihre phonetische Beschreibung umgesetzt. Diese phonetische Beschreibung und die korrespondierenden Sprachsignale werden während der Trainingsphase des Spracherkennungssystems dem Spracherkennungssystem zugeführt. Das Spracherkennungssystem lernt hierdurch den Zielwortschatz. Da der Zielwortschatz von

einer großen Anzahl von Sprechern gesprochen worden ist, ist das Spracherkennungssystem unabhängig von einem einzelnen bestimmten Sprecher.

Das Erstellen einer spezielle Applikation mit einem vorbestimmten Zielwortschatz und das Sprechen durch mehrere Sprecher, so dass eine sprecherunabhängige Sprachdatenbank erzeugt wird, dauert in der Regel zwischen zwei bis sechs Monaten. Das Erstellen derartiger anwendungsspezifischer Sprachdatenbanken verursacht den größten Kostenfaktor beim Anpassen eines bestehenden Spracherkennungssystems an eine bestimmte Applikation. Es besteht deshalb ein erheblicher Bedarf nach einem Verfahren, mit welchen kostengünstig und schnell eine Sprachdatenbank zum Trainieren eines sprecherunabhängigen Spracherkennungssystems erstellt werden kann.

Der Erfindung liegt deshalb die Aufgabe zugrunde, ein Verfahren zum Erzeugen einer Sprachdatenbank für einen Zielwortschatz zum Trainieren eines Spracherkennungssystems zu schaffen, mit welchem schneller und vor allem kostengünstiger die Sprachdatenbank erzeugt werden kann.

Die Aufgabe wird durch ein Verfahren mit den Merkmalen des Anspruchs 1 gelöst. Vorteilhafte Ausgestaltungen der Erfindung sind in den Unteransprüchen angegeben.

Gemäß dem erfindungsgemäßen Verfahren zum Erzeugen einer Sprachdatenbank für einen Zielwortschatz zum Trainieren eines Spracherkennungssystems werden
die Worte des Zielwortschatzes in eine phonetische Beschreibung umgesetzt, so dass die einzelnen Worte durch Phoneme dargestellt werden, und
aus einem oder mehreren Phonen zusammengesetzte Segmente eines gesprochenen, vom Zielwortschatz unabhängigen Trainingstextes, werden zu Wörtern des Zielwortschatzes entsprechend der phonetischen Beschreibung konkateniert bzw. zusammengesetzt.

Mit dem erfindungsgemäßen Verfahren werden Segmente eines vom Zielwortschatz unabhängigen Trainingstextes zu den Wörtern des Zielwortschatzes zusammengesetzt. Der Trainingstext kann somit ein beliebiger bereits vorab aufgezeichneter und in einzelne Sprachsegmente segmentierter Text sein. Zum Erzeugen der Sprachdatenbank ist es deshalb nicht notwendig, jedes Mal einen den Zielwortschatz enthaltenden Trainingstext zu erstellen und aufzuzeichnen. Es ist vielmehr möglich, vorhandene Sprachdatenbanken mit allgemeinen Wortschätzen zu verwenden. Die Wörter dieser vorhandenen Sprachdatenbanken werden vorab segmentiert. Diese Segmentierung kann manuell oder automatisch erfolgen. Grundsätzlich genügt es, dass für jede Sprache eine derartig segmentierte Datenbank lediglich ein einziges Mal vorliegt. Ausgehend von dieser Datenbank wird mit dem erfindungsgemäßen Verfahren eine für eine Applikation spezifische Sprachdatenbank erzeugt. Ein erneutes Sprechen eines Trainingstextes ist somit nicht notwendig.

Mit dem erfindungsgemäßen Verfahren kann schnell und kostengünstig eine zum Trainieren eines sprecherunabhängigen Spracherkennungssystems geeignete Sprachdatenbank erzeugt werden, wobei es nicht notwendig ist, dass spezielle Trainingstexte aufgezeichnet werden, wodurch die Kosten im Vergleich zu den bekannten Methoden zum Erstellen derartiger Sprachdatenbanken drastisch vermindert werden.

Vorzugsweise werden die Wörter des Zielwortschatzes aus möglichst langen Segmenten zusammengesetzt. Sollte dies nicht möglich sein, müssen den einzelnen Phonemen der Wörter relativ kurze Segmente mit jeweils einem einzigen Phon zugeordnet und zu dem entsprechenden Wort konkateniert werden. Dies erfolgt vorzugsweise unter Berücksichtigung des Kontextes, in dem die jeweiligen Phoneme der Wörter und der Phone des Trainingstextes angeordnet sind.

Nach einer weiteren bevorzugten Ausführungsform werden konkatenierte Segmente an ihren Grenzen zwischen zwei benachbarten Segmenten geglättet.

Die Segmente können in Form von elektrischen Sprachsignalen oder als Merkmalsvektoren vorliegen. Letztere Darstellungsform ist vorteilhaft, da der Datenumfang der Merkmalsvektoren deutlich geringer als der der elektrischen Sprachsignale ist.

Die Erfindung wird nachfolgend beispielhaft anhand der beiliegenden Zeichnungen näher erläutert. In denen zeigen schematisch:

Fig. 1 ein Verfahren zum Aufbereiten eines aufgezeichneten Trainingstextes für das erfindungsgemäße Verfahren,

Fig. 2 einen allgemeinen Überblick über die Abläufe beim erfindungsgemäßen Verfahren zum Erzeugen einer Sprachdatenbank in einem Flussdiagramm,

Fig. 3 das Verfahren zum Konkatenieren der Wörter des Zielwortschatzes aus Sprachsegmenten in einem Flussdiagramm, und

Fig. 4 ein Computersystem zum Ausführen des erfindungsgemäßen Verfahren in einem Blockschaltbild.

Das erfindungsgemäße Verfahren betrifft im Allgemeinen das Konkatenieren bzw. Zusammensetzen von Wörtern eines Zielwortschatzes aus Segmenten eines gesprochenen Textes.

Der Zielwortschatz ist phonetisch zu beschreiben, das heißt, dass die einzelnen Worte durch Phoneme dargestellt sind. Die Segmente sind aus einzelnen Phonen zusammengesetzt.

Im Sinne der vorliegenden Beschreibung der Erfindung ist ein Phonem die kleinste bedeutungsunterscheidende, aber nicht

selbst bedeutungstragende sprachliche Einheit (z.B. b in Bein im Unterschied zu p in Pein). Ein Phon ist hingegen der ausgesprochene Laut eines Phonems. Phoneme werden in einer Lautschrift dargestellt, wobei jeder „Buchstabe" der Lautschrift ein Phonem darstellt. Phone werden durch physikalische Größen dargestellt, die den ausgesprochenen Laut an sich wiedergeben. Diese physikalischen Größen können elektrische Sprachsignale sein, die an einem Lautsprecher in entsprechende akustische, den Laut darstellende Signale gewandelt werden können. Phone können jedoch auch durch sogenannte Merkmalsvektoren dargestellt werden. Merkmalsvektoren umfassen Koeffizienten, die das entsprechende Sprachsignal zu einem bestimmten Zeitpunkt wiedergeben. Derartige Koeffizienten werden durch Abtasten des Sprachsignals in vorbestimmten Zeitabständen erhalten. Typische Zeitabstände sind 10 ms bis 25 ms. Bekannte Koeffizienten sind die ACF-Koeffizienten (Auto-Correlation-Function) und die LPCC-Koeffizienten (Linear-Predictive-Cepstral Coeffizient).

Die obigen Erläuterungen können dahingehend kurz zusammengefasst werden, dass Phoneme die symbolische Beschreibung einzelner Laute und Phone die physikalische Beschreibung der Laute sind.

Nachfolgend wird anhand von Fig. 1 ein Verfahren zum Aufbereiten eines Trainingstextes in Sprachsegmente erläutert, wobei die Sprachsegmente ein oder mehrere Phone umfassen.

Das Verfahren beginnt im Schritt S1. Im Schritt S2 werden von mehreren Sprechern ein oder mehrere Trainingstexte gesprochen und elektronisch aufgezeichnet.

Die elektronisch aufgezeichneten Trainingstexte werden im Schritt S3 zur Datenreduktion in Merkmalsvektoren umgesetzt.

Die derart gespeicherte Sprachaufzeichnung wird im Schritt S4 in Segmente aufgeteilt, wobei die einzelnen Segmente jeweils

ein einziges Phon umfassen. Diese Segmentierung wird in der Regel automatisch durchgeführt. Die Sprachaufzeichnung kann jedoch bereits vor der Umsetzung in Merkmalsvektoren manuell von einem Sprachexperten vorsegmentiert werden.

Diese jeweils ein einziges Phon umfassenden Segmente werden im Schritt S5 statistisch erfasst, wobei typische Laute und Lautfolgen statistisch ausgewertet und festgehalten werden. Diese statistischen Informationen über die Lautfolgen ergeben in Verbindung mit den Segmenten, die jeweils nur ein einziges Phon enthalten, eine Darstellung der im Trainingstext enthaltenen Segmente mit mehreren Phonen wieder. Hierdurch stehen für die weitere Auswertung nicht nur Segmente mit einem einzigen Phon, sondern auch längere Segmente mit zumindest zwei Phonen zur Verfügung.

Im Schritt S5 wird vorzugsweise eine Energie-Normierung der einzelnen Segmente ausgeführt, da die unterschiedlichen Sprecher üblicherweise mit einer unterschiedlichen Lautstärke sprechen, so dass die einzelnen Segmente unterschiedlicher Sprecher nicht miteinander vergleichbar und oftmals auch nicht zu einem neuen Wort zusammensetzbar sind.

Dieses Verfahren zum Aufbereiten der Segmente wird im Schritt S6 beendet.

Mit dem in Fig. 1 gezeigten Verfahren zum Aufbereiten eines Trainingstextes wird eine Segmentdatenbank erstellt. Grundsätzlich genügt es, dass für jede Sprache, für die das erfindungsgemäße Verfahren angewendet werden soll, lediglich eine einzige Segmentdatenbank erstellt wird. Als Trainingstexte werden allgemeine Texte verwendet, die für die wichtigsten Sprachen bereits als Datenbank in Form von beispielsweise einer ASCII-Datei für die Texte und in Form von Sprachsignalen in großem Umfang existieren.

In Fig. 2 ist der allgemeine Ablauf des erfindungsgemäßen Verfahrens zum Erzeugen einer Sprachdatenbank für einen vorgegebenen Zielwortschatz in Form eines Flußdiagrammes dargestellt.

Der Zielwortschatz liegt als Textdatei (z.B. ASCII-Datei) vor. Der Zielwortschatz umfasst die für die beabsichtigte Applikation notwendigen Wörter. Solche Zielwortschätze können beispielsweise nur wenige Wörter (z.B. 20 bis 50 Wörter) umfassen, die beispielsweise zum Ansteuern eines bestimmten Gerätes notwendig sind. Es ist jedoch auch möglich, noch kleinere mit sogar nur einem einzigen Wort oder auch größere Zielwortschätze vorzusehen, die beispielsweise einige tausend Wörter umfassen.

Das Verfahren zum Erzeugen einer Sprachdatenbank beginnt mit dem Schritt S7. Im Schritt S8 werden die Wörter des Zielwortschatzes in ihre phonetische Beschreibung umgesetzt. Hierzu sind regelbasierte Verfahren bekannt, die automatisch eine derartige Umsetzung vornehmen. Grundsätzlich ist es auch möglich, statistische Verfahren zu verwenden. Neuere Verfahren zum Umsetzen einer Textdatei in ihre phonetische Schreibweise beruhen auf neuronalen Netzwerken.

Im darauffolgenden Schritt S9 werden die Segmente des Trainingstextes zu den einzelnen Wörtern des Zielwortschatzes konkateniert. Hierbei werden Segmente, deren Phone den Phonemen der Wörter des Zielwortschatzes entsprechen zu den entsprechenden Wörtern zusammengesetzt bzw. konkateniert.

Sind alle Wörter des Zielwortschatzes konkateniert, kann im Schritt S10 eine Nachbearbeitung durchgeführt werden. Hierbei wird bspw. eine Datenreduktion durchgeführt, falls die konkatenierten Wörter als Sprachsignal vorliegen.

Im Schritt S11 ist das Verfahren beendet.

In Fig. 3 sind in einem Flussdiagramm die einzelnen beim Konkatenieren auszuführenden Verfahrensschritte dargestellt.

Dieser Konkateniervorgang beginnt mit dem Schritt S12. Zunächst wird im Schritt S13 ein Wort des Zielwortschatzes ausgewählt, das zu Konkatenieren ist.

Im Schritt S14 wird versucht, das ausgewählte Wort mittels einem einzigen oder wenigen langen Segmenten zusammen zu setzen. Hierbei werden aus der Segmentdatenbank Segmente ausgewählt, deren Phonemzuordnung mit den Phonemen des zu konkatenierenden Wortes übereinstimmt.

Im Schritt S15 wird abgefragt, ob das Wort aus den langen Segmenten erfolgreich konkateniert werden konnte. Ist das Ergebnis dieser Abfrage nein, so bedeutet dies, dass keine geeigneten langen Segmente in der Segmentdatenbank vorhanden sind, aus welchen das Wort zusammengesetzt werden kann. Der Verfahrensablauf geht deshalb auf den Schritt S16 über, bei dem das Wort aus einzelnen Phonemen unter Berücksichtigung des entsprechenden Kontextes konkateniert wird. Hierbei werden Segmente mit einem einzigen Phon den korrespondierenden Phonemen des zu konkatenierenden Wortes zugeordnet, wobei jedoch nur Phone verwendet werden, deren benachbarte Phone im Trainingstext den zu dem jeweiligen Phonem benachbarten Phonemen im zu konkatenierenden Wort entsprechen. Wird z.B. das Phon „f" dem Phonem „f" im Wort „Anfang" zugeordnet, so wird ein Segment mit dem Phon „f" aus dem Trainingstext gewählt, das im Trainingstext zwischen den Phonen „n" und „a" angeordnet ist. Der Kontext „nfa" des Segmentes „f" stimmt somit mit dem Kontext des Phonems „f" des Wortes aus dem Zielwortschatz überein.

Im Schritt S17 wird geprüft, ob das zu konkatenierende Wort vollständig konkateniert werden konnte. Ergibt diese Überprüfung ein „nein", so geht der Verfahrensablauf auf den Schritt S18 über. Im Schritt S18 werden für diejenigen Phoneme, denen

noch keine Segmente zugeordnet werden konnten, Segmente ausgewählt, deren Phon mit dem entsprechenden Phonem möglichst übereinstimmt und deren Kontext möglichst ähnlich ist. Sind keine Segmente mit Phonen, die den Phonemen unmittelbar entsprechen, vorhanden, werden solche Segmente ausgewählt, deren Phone den Phonemen möglichst ähnlich sind.

Die Ähnlichkeit der Kontexte bzw. der Phone zu den einzelnen Phonemen wird nach vorbestimmten Regeln beurteilt. Diese Regeln können z.B. als Listen in einer speziellen Ähnlichkeitsdatenbank abgespeichert sein, wobei zu jedem Phonem eine Liste weiterer Phoneme gespeichert ist, und die weiteren Phoneme mit abnehmender Ähnlichkeit sortiert sind. Zu dem Phonem „p" ist z.B. folgende Liste mit „b, d, t, ..." gespeichert. Dies bedeutet, dass das Phonem „b" am ähnlichsten zu dem Phonem „p" ist und das Phonem „d" das zweitähnlichste Phonem ist. Die Ähnlichkeitsdatenbank kann auch Kontexte mit zwei oder mehreren Phonemen umfassen. Zum Kontext „_a_s" wird z.B. die Liste „_a_f, _a_x, ..." abgespeichert. Dies bedeutet, dass der Kontext „_a_f" am ähnlichsten zu „_a_s" ist. Die Reihenfolge der gespeicherten ähnlichen Phoneme kann sich je nach der Definition des Kriteriums der „Ähnlichkeit" unterscheiden. Die oben verwendete Notation „_a_s" ist eine firmeninterne Notation und bedeutet:

_a_s: Phonem a mit Rechtskontext s
_a_x: Phonem a mit Rechtskontext x
t_a_: Phonem a mit Linkskontext t
p_a_: Phonem a mit Linkskontext p usw.

Anstelle derartiger Listen oder in Ergänzung zu diesen Listen können auch allgemeine Regeln zum Vergleich von ähnlichen Kontexten vorgesehen sein. So können in einem Kontext z.B. Plosive oder Frikative grundsätzlich als sehr ähnlich beurteilt werden.

Nach dem Zuordnen der ähnlichsten Segmente zu den entsprechenden Phonemen des zu konkatenierenden Wortes geht der Ver-

fahrensablauf auf den Schritt S19 über. Sollte sich bei den abfragenden Schritten S15 und S17 ergeben, dass die Konkatenierung erfolgreich ausgeführt worden ist, so geht auch hier der Verfahrensablauf unmittelbar auf den Schritt S19 über.

Im Schritt S19 erfolgt die Endbearbeitung der einzelnen konkatenierten Wörter. Diese Endbearbeitung kann folgende Teilschritte einzeln oder in Kombination umfassen:

- Am Anfang und am Ende eines soeben konkatenierten Wortes wird eine für den Anfang und das Ende des Wortes typische Geräuschsequenz angefügt.

- Die einzelnen Segmente in den Wörtern werden normiert. Dies ist insbesondere zweckmäßig, wenn eine Segmentdatenbank mit nicht-normierten Segmenten verwendet wird.

- Die Übergänge an den Grenzen zwischen zwei benachbarten Segmenten werden geglättet, wobei die erste und die zweite Ableitung des Sprachsignals oder der Koeffizienten der Merkmalsvektoren an der Übergangsstelle möglichst 0 beträgt.

Im Schritt S20 wird geprüft, ob noch ein weiteres Wort des Zielwortschatzes zu konkatenieren ist. Ergibt die Abfrage ein ja, so geht der Verfahrensablauf auf den Schritt S13 über, wohingegen das Verfahren im Schritt S21 beendet wird, falls die Abfrage ein nein ergibt.

Die mit dem erfindungsgemäßen Verfahren konkatenierten Wörter des Zielwortschatzes stellen eine Sprachdatenbank dar, mit der ein Spracherkennungssystem auf den Zielwortschatz trainiert werden kann. Zum Erstellen dieser Sprachdatenbank ist es nicht notwendig, dass spezielle den Zielwortschatz enthaltende Trainingstexte erstellt werden, die von Sprechern gesprochen und aufgezeichnet werden müssen. Vielmehr können durch das erfindungsgemäße Verfahren ein allgemeiner Trainingstext, der einmal von einem oder mehreren Sprechern ge-

sprochen worden ist, und entsprechend segmentiert worden ist, zur Erzeugung einer Sprachdatenbank für einen speziellen Zielwortschatz ausgewertet werden. Dies bedeutet einen erheblichen Zeitgewinn und eine enorme Kosteneinsparung bei der Erzeugung einer Sprachdatenbank für einen speziellen Zielwortschatz.

Mit einem sehr vereinfachten Prototypen des erfindungsgemäßen Verfahrens ist ein Zielwortschatz mit zehn Wörtern konkateniert worden, wobei lediglich Segmente mit einem oder zwei Phonen berücksichtigt worden sind. Bei diesem Prototypen wurde weder eine Normierung vorgenommen, noch die Übergänge zwischen benachbarten Segmenten geglättet. Zudem beruhte die Segmentdatenbank auf einem Trainingstext mit nur 60 unterschiedlichen Wörtern.

Trotz dieser sehr geringen Datenmenge und des stark vereinfachten Verfahrens ist eine Erkennungsrate von ca. 80% erzielt worden.

Die Erfindung ist oben anhand eines Ausführungsbeispiels näher erläutert worden. Sie ist jedoch nicht auf das konkrete Ausführungsbeispiel beschränkt. So ist es z.B. im Rahmen der Erfindung möglich, für jedes Phonem eines zu konkatenierenden Wortes mehrere ähnliche Segmente aus der Segmentdatenbank auszuwählen und diese dann aufgrund ihrer Ähnlichkeit zum Phonem bzw. zum Kontext der aus zwei, drei, vier oder mehreren Phonemen bestehen kann, zu bewerten. Das ähnlichste Segment wird ausgewählt. Es ist jedoch auch möglich, eine Gruppe ähnlicher Segmente auszuwählen und anstelle ein einzelnes Segment zu bestimmen, das dem Phonem zugeordnet wird, aus dieser Gruppe von Segmenten ein mittleres Segment zu berechnen, das dem Phonem zugeordnet wird. Dies ist insbesondere dann zweckmäßig, wenn die Phone der Segmente durch Merkmalsvektoren beschrieben werden, die gemittelt werden können. Anstelle einer Mittlung der mehreren Segmente kann auch ein Segment bestimmt werden, dessen Abstand (Vektorabstand der

Merkmalsvektoren) zu den ausgewählten Segmenten am geringsten ist.

Das erfindungsgemäße Verfahren kann als Computerprogramm realisiert werden, das selbstständig auf einem Computer zum Erzeugen einer Sprachdatenbank aus einer Segmentdatenbank ablaufen kann. Es stellt somit ein automatisch ausführbares Verfahren dar.

Das Computerprogramm kann auf elektrisch lesbaren Datenträgern gespeichert werden und so auf andere Computersysteme übertragen werden.

Ein zur Anwendung des erfindungsgemäßen Verfahrens geeignetes Computersystem ist in Fig. 4 gezeigt. Das Computersystem 1 weist einen internen Bus 2 auf, der mit einem Speicherbereich 3, einer zentralen Prozessoreinheit 4 und einem Interface 5 verbunden ist. Das Interface 5 stellt über eine Datenleitung 6 eine Datenverbindung zu weiteren Computersystemen her. An dem internen Bus 2 sind ferner eine akustische Eingabeeinheit 7, eine grafische Ausgabeeinheit 8 und eine Eingabeeinheit 9 angeschlossen. Die akustische Eingabeeinheit 7 ist mit einem Lautsprecher 10, die grafische Ausgabeeinheit 8 mit einem Bildschirm 11 und die Eingabeeinheit 9 mit einer Tastatur 12 verbunden. An dem Computersystem 1 kann beispielsweise über die Datenleitung 6 und das Interface 5 ein Zielwortschatz übertragen werden, der im Speicherbereich 3 abgespeichert wird. Der Speicherbereich 3 ist in mehrere Bereiche unterteilt, in denen der Zielwortschatz, das Programm zum Ausführen des erfindungsgemäßen Verfahrens und weitere Anwendungs- und Hilfsprogramme gespeichert sind. Mit dem erfindungsgemäßen Verfahren wird eine Sprachdatenbank zum Zielwortschatz erstellt. Diese Sprachdatenbank wird dann zum Trainieren eines Spracherkennungssystems verwendet. Das Spracherkennungssystem kann eingehende Audiodateien in Text-Dateien automatisch umsetzen. Die Audiodateien können durch Sprechen eines Testes in das Mikrofon 10 erzeugt werden.

Patentansprüche

1. Verfahren zum Erzeugen einer Sprachdatenbank für einen Zielwortschatz zum Trainieren eines Spracherkennungssystems, wobei

die Worte des Zielwortschatzes in eine phonetische Beschreibung umgesetzt werden (S8), so dass die einzelnen Worte durch Phoneme dargestellt werden, und

aus einem oder mehreren Phonen zusammengesetzte Segmente eines gesprochenen, vom Zielwortschatz unabhängigen Trainingstextes, zu Wörtern des Zielwortschatzes entsprechend der phonetischen Beschreibung konkateniert werden (S9).

2. Verfahren nach Anspruch 1,
d a d u r c h g e k e n n z e i c h n e t,
dass möglichst lange Segmente ausgewählt werden, aus welchen die Wörter konkateniert werden.

3. Verfahren nach Anspruch 2,
d a d u r c h g e k e n n z e i c h n e t,
dass zu den einzelnen Segmenten Kontextinformationen von einem oder mehrerer benachbarter Phone im Trainingstext gespeichert sind (S5), und falls nicht alle Phoneme eines Wortes aus Segmenten mit mindestens zwei Phonen konkatenierbar sind, Segmente mit jeweils einem einzigen Phon ausgewählt werden, deren Phone den nicht aus längeren Segmenten konkatenierbaren Phonemen im zu konkatenierenden Wort entsprechen und deren Kontextinformationen mit den Kontexten dieser Phoneme im zum konkatenierenden Wort übereinstimmen (S17).

4. Verfahren nach Anspruch 3,
d a d u r c h g e k e n n z e i c h n e t,
dass beim Konkatenieren von Segmenten mit einzelnen Phonen, falls keine Segmente deren Kontextinformation mit den Kontexten der Phoneme des zu konkatenierenden Wortes überein-

stimmen, vorhanden sind, Segmente mit jeweils einem einzigen Phon ausgewählt werden, deren Phone den nicht aus längeren Segmenten konkatenierbaren Phonemen im zu konkatenierenden Wort entsprechen und deren Kontextinformationen mit den Kontexten dieser Phoneme im zum konkatenierenden Wort möglichst ähnlich sind (S18).

5. Verfahren nach einem der Ansprüche 1 bis 4
d a d u r c h g e k e n n z e i c h n e t,
dass die zu konkatenierten Segmente an den Grenzen zwischen zwei benachbarten Segmenten geglättet werden (S19).

6. Verfahren nach einem der Ansprüche 1 bis 5,
d a d u r c h g e k e n n z e i c h n e t,
dass die einzelnen Segmente vor dem Konkatenieren energienormiert werden (S19).

7. Verfahren nach einem der Ansprüche 1 bis 6,
d a d u r c h g e k e n n z e i c h n e t,
dass die Segmente in Form von elektrischen Sprachsignalen vorliegen.

8. Verfahren nach einem der Ansprüche 1 bis 6,
d a d u r c h g e k e n n z e i c h n e t,
dass die Segmente durch Merkmalsvektoren dargestellt sind.

9. Verfahren nach Anspruch 8,
d a d u r c h g e k e n n z e i c h n e t,
dass falls beim Auswählen der Segmente vor dem Konkatenieren mehrere dem Phonem bzw. den Phonemen des zu konkatenierenden Wortes zuordbare Segmente vorhanden sind, ein Segment entweder durch Mitteln der Merkmalsvektoren der zuordbaren Segmente oder dasjenige Segment bestimmt wird, dessen Merkmalsvektor den geringsten mittleren Abstand zu den zuordbaren Segmenten besitzt.

10. Verfahren nach einem der Ansprüche 1 bis 9,
d a d u r c h g e k e n n z e i c h n e t,
dass die zu Worten konkatenierten Segmente einer Datenreduktion unterzogen werden.

11. Vorrichtung zum Erzeugen einer Sprachdatenbank für einen Zielwortschatz zum Trainieren eines Spracherkennungssystems, mit

einem Computersystem (1), das einen Speicherbereich (3) aufweist, in dem ein Computerprogramm zum Ausführen eines Verfahrens nach einem der Ansprüche 1 bis 10 gespeichert ist.

FIG 1

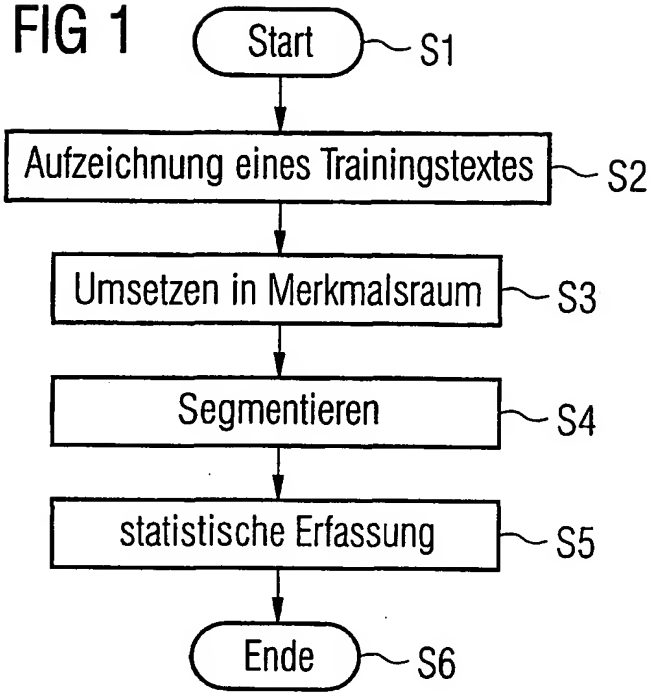

Start
S1
Aufzeichnung eines Trainingstextes
S2
Umsetzen in Merkmalsraum
S3
Segmentieren
S4
statistische Erfassung
S5
Ende
S6

FIG 2

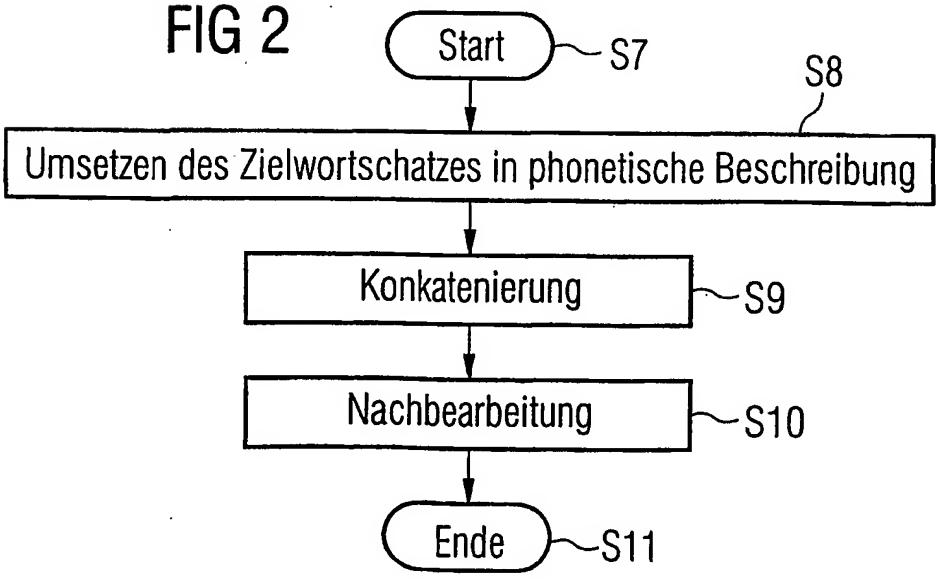
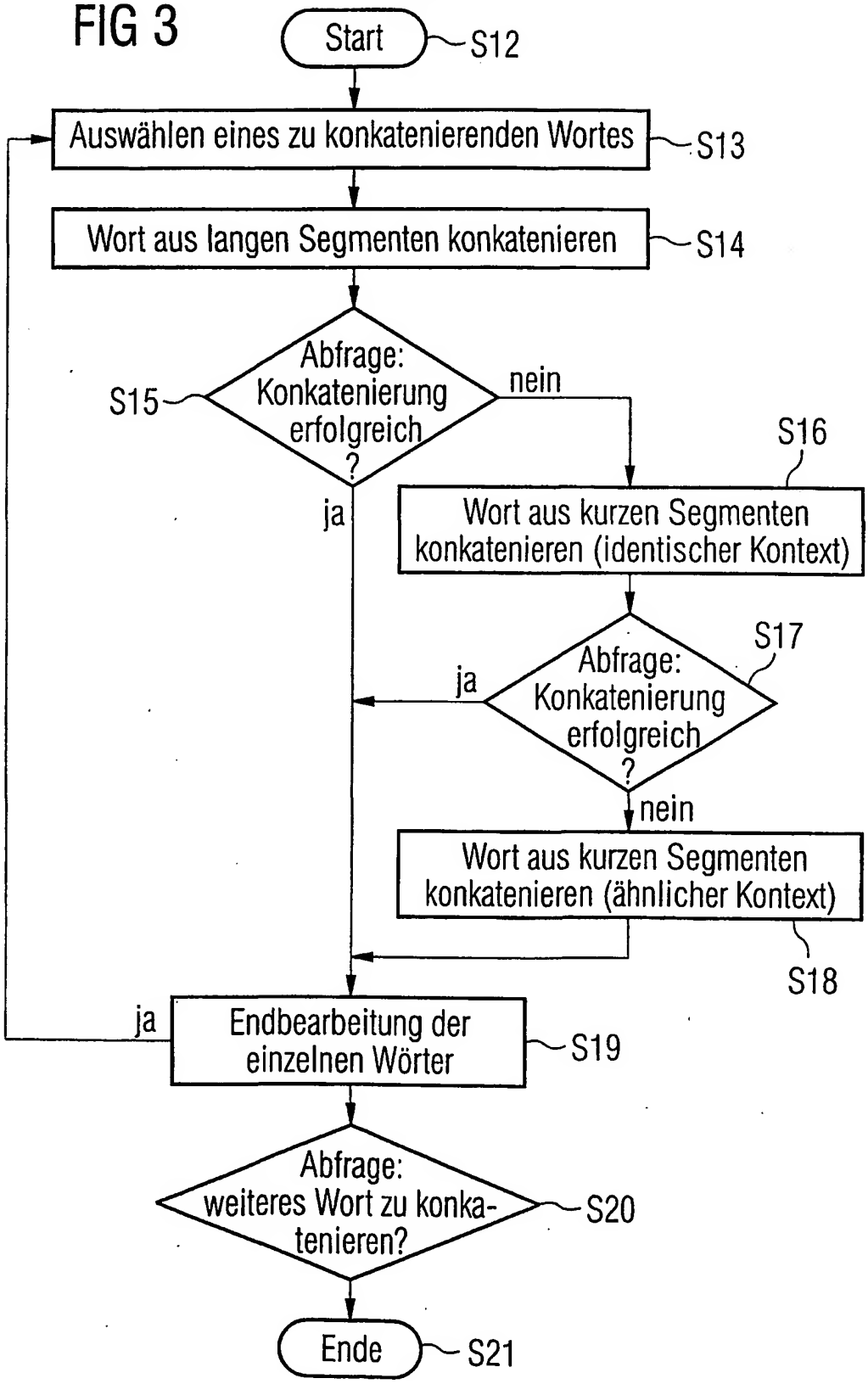

Start
S7
S8
Umsetzen des Zielwortschatzes in phonetische Beschreibung
Konkatenierung
S9
Nachbearbeitung
S10
Ende
S11

FIG 3


Start
S12
Auswählen eines zu konkatenierenden Wortes
S13
Wort aus langen Segmenten konkatenieren
S14
Abfrage: Konkatenierung erfolgreich ?
S15
nein
ja
S16
Wort aus kurzen Segmenten konkatenieren (identischer Kontext)
Abfrage: Konkatenierung erfolgreich ?
S17
ja
nein
Wort aus kurzen Segmenten konkatenieren (ähnlicher Kontext)
S18
ja
Endbearbeitung der einzelnen Wörter
S19
Abfrage: weiteres Wort zu konkatenieren?
S20
Ende
S21

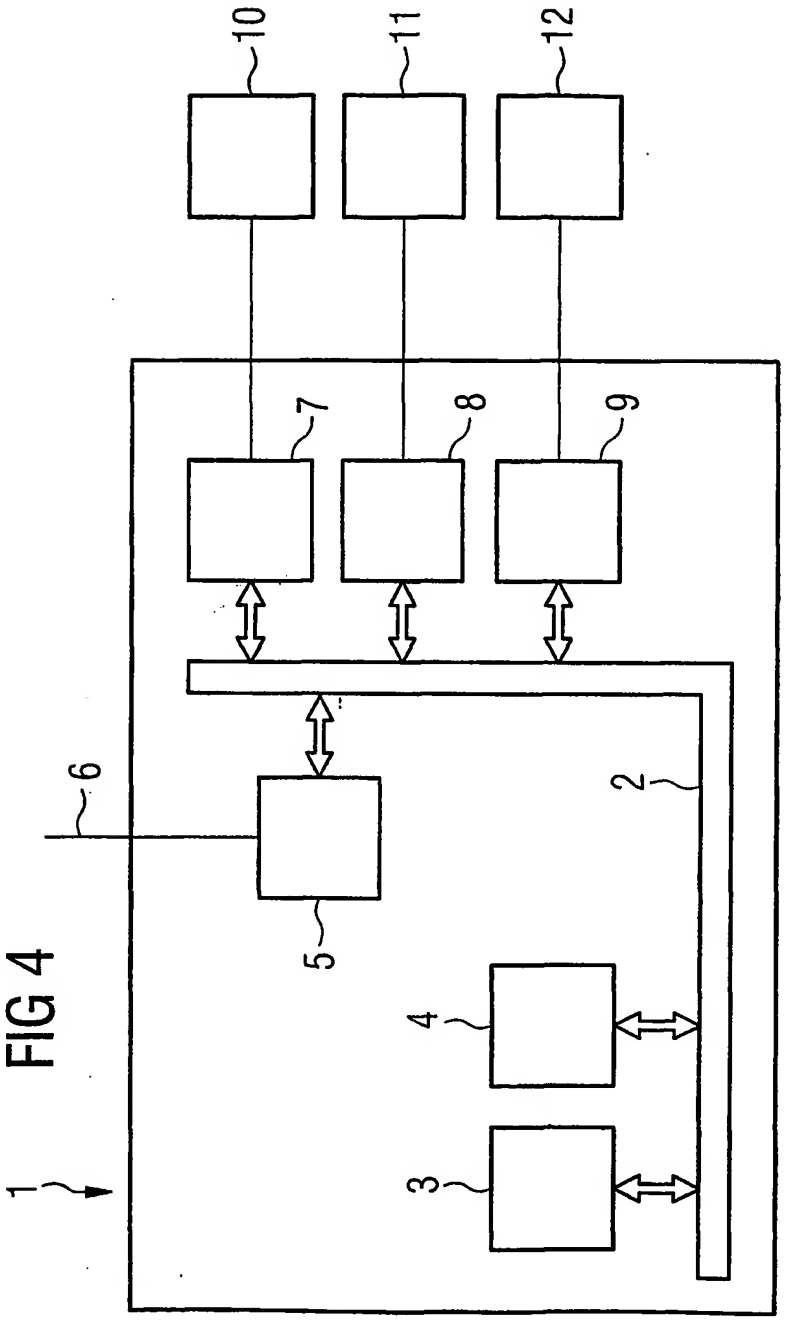


FIG 4

INTERNATIONAL SEARCH REPORT

tional Application No
PCT/DE 01/01546

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, COMPENDEX, PAJ, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 579 436 A (JUANG BIING-HWANG ET AL) 26 November 1996 (1996-11-26) column 3, line 57 -column 4, line 4; claim 1	1,11

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 July 2001	13/08/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	De Vos, L

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Ir tional Application No
I .,/DE 01/01546

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	HUNT ANDREW J ET AL: "Unit selection in a concatenative speech synthesis system using a large speech database" PROCEEDINGS OF THE 1996 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP. PART 1 (OF 6);ATLANTA, GA, USA MAY 7-10 1996, vol. 1, 1996, pages 373-376, XP002172898 ICASSP IEEE Int Conf Acoust Speech Signal Process Proc;ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 1996 IEEE, Piscataway, NJ, USA abstract	1,11
A	US 5 850 627 A (STURTEVANT DEAN G ET AL) 15 December 1998 (1998-12-15) column 56, line 7 -column 57, line 20; figure 72	1,11
P,A	BENTEZ M C ET AL: "Different confidence measures for word verification in speech recognition" SPEECH COMMUNICATION,ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM,NL, vol. 32, no. 1-2, September 2000 (2000-09), pages 79-94, XP004216247 ISSN: 0167-6393 page 83, left-hand column, line 8 - line 10	1,11

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

Ional Application No
PCT/DE 01/01546

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5579436 A	26-11-1996	CA 2089903 A,C	03-09-1993
		DE 69322894 D	18-02-1999
		DE 69322894 T	29-07-1999
		EP 0559349 A	08-09-1993
		ES 2128390 T	16-05-1999
		JP 3053711 B	19-06-2000
		JP 6012093 A	21-01-1994
US 5850627 A	15-12-1998	US 5428707 A	27-06-1995
		US 6073097 A	06-06-2000
		US 5920836 A	06-07-1999
		US 5920837 A	06-07-1999
		US 6101468 A	08-08-2000
		US 5983179 A	09-11-1999
		US 5909666 A	01-06-1999
		US 5915236 A	22-06-1999
		US 5960394 A	28-09-1999
		US 6092043 A	18-07-2000

INTERNATIONALER RECHERCHENBERICHT

I	tionales Aktenzeichen
	rct/DE 01/01546

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G10L15/06

Nach der internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G10L

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, COMPENDEX, PAJ, WPI Data, INSPEC

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	US 5 579 436 A (JUANG BIING-HWANG ET AL) 26. November 1996 (1996-11-26) Spalte 3, Zeile 57 -Spalte 4, Zeile 4; Anspruch 1 --- -/--	1,11

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

[X] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :
'A' Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
'E' älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
'L' Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
'O' Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
'P' Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
'T' Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
'X' Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
'Y' Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
'&' Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
24. Juli 2001	13/08/2001

Name und Postanschrift der internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	De Vos, L

Formblatt PCT/ISA/210 (Blatt 2) (Juli 1992)

INTERNATIONALER RECHERCHENBERICHT

	itionales Aktenzeichen
	rct/DE 01/01546

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	HUNT ANDREW J ET AL: "Unit selection in a concatenative speech synthesis system using a large speech database" PROCEEDINGS OF THE 1996 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP. PART 1 (OF 6);ATLANTA, GA, USA MAY 7-10 1996, Bd. 1, 1996, Seiten 373-376, XP002172898 ICASSP IEEE Int Conf Acoust Speech Signal Process Proc;ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 1996 IEEE, Piscataway, NJ, USA Zusammenfassung	1,11
A	US 5 850 627 A (STURTEVANT DEAN G ET AL) 15. Dezember 1998 (1998-12-15) Spalte 56, Zeile 7 -Spalte 57, Zeile 20; Abbildung 72	1,11
P,A	BENTEZ M C ET AL: "Different confidence measures for word verification in speech recognition" SPEECH COMMUNICATION,ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM,NL, Bd. 32, Nr. 1-2, September 2000 (2000-09), Seiten 79-94, XP004216247 ISSN: 0167-6393 Seite 83, linke Spalte, Zeile 8 - Zeile 10	1,11

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentli[illegible]en, die zur selben Patentfamilie gehören

[illegible]ionales Aktenzeichen
PCT/DE 01/01546

Im Recherchenbericht angeführtes Patentdokument		Datum der Veröffentlichung	Mitglied(er) der Patentfamilie		Datum der Veröffentlichung
US 5579436	A	26-11-1996	CA	2089903 A,C	03-09-1993
			DE	69322894 D	18-02-1999
			DE	69322894 T	29-07-1999
			EP	0559349 A	08-09-1993
			ES	2128390 T	16-05-1999
			JP	3053711 B	19-06-2000
			JP	6012093 A	21-01-1994
US 5850627	A	15-12-1998	US	5428707 A	27-06-1995
			US	6073097 A	06-06-2000
			US	5920836 A	06-07-1999
			US	5920837 A	06-07-1999
			US	6101468 A	08-08-2000
			US	5983179 A	09-11-1999
			US	5909666 A	01-06-1999
			US	5915236 A	22-06-1999
			US	5960394 A	28-09-1999
			US	6092043 A	18-07-2000